DisSim: A Discourse-Aware Syntactic Text Simplification Framework for English and German





Christina Niklaus, Matthias Cetto, André Freitas and Siegfried Handschuh

PROBLEM

SEMANTIC

HIERARCHY

set of hierarchically ordered and semantically

propositions that present a simplified syntax.

to produce a disconnected sequence of output

CONTRAST

sentences that are hard to interpret.

A fluoroscopic study is

typically the next step

in management.

This fluoroscopic study

is known as an upper

gastrointestinal series.

ENABLEMENT

BACKGROUND

no more rule matches.

logical structure of

utterances that present a

simplified syntax

The usage of barium can impede

surgical revision.

interconnected sentences in the form of minimal

DisSim breaks down a complex source sentence into a

By taking into account discourse-level aspects, we avoid

BACKGROUNI

Caution with non water

soluble contrast is

mandatory.

CONDITION

Volvulus is suspected.

The usage of barium can lead to

increased post operative

Complex source sentence:

"A fluoroscopic study known as an upper gastrointestinal series is typically the next step in management, although if volvulus is suspected, caution with non water soluble contrast is mandatory as the usage of barium can impede surgical revision and lead to increased post operative complications."

Minimality:

improves the performance of SOTA Open IE approaches in terms of precision (up to 346%) and recall (up to 52%), i.e. leads to a lower information loss and a higher accuracy of the extracted tuples.

Recall

+52%

+40%

+8%

-1%

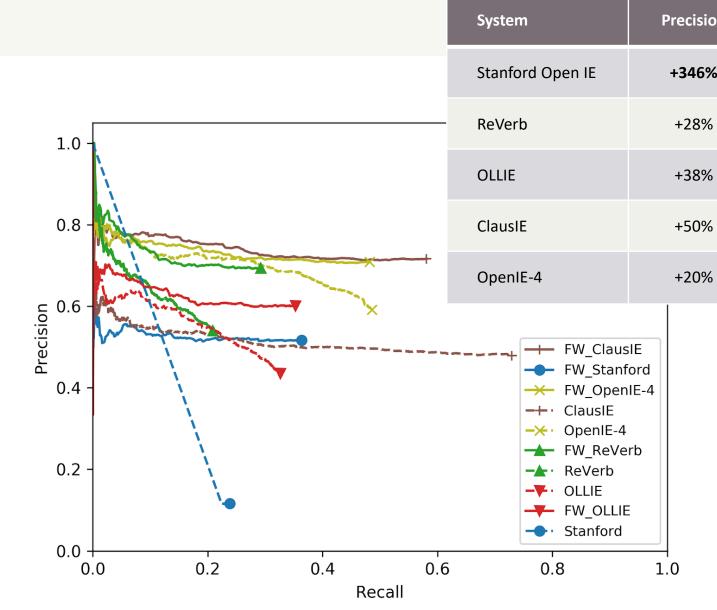
AUC

+20%

+15%

propositions

Minimal



Lightweight semantic representation can be used to **facilitate and improve** a variety of AI tasks.

Open IE

Supervised-OIE (alone). (1) (A fluoroscopic study; known; as an upper gastrointestinal series) (2) (caution with non water soluble contrast; is; mandatory as the usage of bariui (3) (as the usage; of barium can impede; surgical revision and lead (4) (; to increased; post operative complications) Supervised-OIE (using DisSim for preprocessing). (5) #1 0 (A fluoroscopic study; is; typically, the next step in management) (This fluoroscopic study; is known; as an upper gastrointestinal series) CONDITION **BACKGROUND #4 BACKGROUND #5 BACKGROUND #6** The usage of barium; can impede; surgical revision) (The usage of barium; can lead; to increased operative complications (The usage of barium; to increased; post operative complications) (11) #7 1 (Volvulus; is suspected;

Semantic **Hierarchy**: enriches the output with important contextual information that is needed for a proper interpretation of complex assertions.

SUBTASK 1

DisSim outperforms the SOTA in structural text simplification: It generates a fine-grained output with a high level of

- grammaticality and meaning preservation. It achieves an improvement of 5%, 4% and 6% with regard to the SAMSA score [3] against the second-best performing approach.
- It can be applied in a domain independent manner.
- The full evaluation methodology and detailed results are reported in [4].

SUBTASK 2

A comparative analysis with the annotations contained in the **RST Discourse Treebank** [5] demonstrates that

- DisSim captures the contextual hierarchy between the split sentences with a precision of 90%, and
- DisSim reaches an average precision of 70% for the classification of the rhetorical relations that hold between them.

APPROACH Recursive transformation process:

35 (English) and 29 (German) hand-crafted grammar rules

Patterns encode syntactic and lexical features that can be derived from a sentence's parse tree

ROOT <<: (S < (NP \$.. (VP < +(VP) (SBAR < , (IN \$+ (|S < (NP \$.. VP |)))))))S < (NP \$.. VP)

Each **rule** specifies:

- How to **split up and rephrase** the input?
- How to set up a **semantic hierarchy**?

Split into Minimal **Propositions**

a sequence of sound, selfcontained utterances, with each of them presenting a minimal semantic unit that cannot be further decomposed into meaningful propositions

Simplification rules encode both the splitting points and rephrasing procedure for disembedding clausal and phrasal components and transforming them into syntactically simplified

A fluoroscopic study ... is typically the next step in

If volvulus is suspected, caution with non water soluble contrast is mandatory as ... operative complications

SUBTASK 1

simpler, more regular structures

stand-alone sentences.

loose arrangement of selfcontained units that lack important contextual information

SUBTASK 2

Establish a **Semantic** Hierarchy

1. Constituency Type Classification: Leaf nodes are recursively simplified in a top-down fashion until

establishes a contextual hierarchy between the split sentences by connecting them with information about their hierarchical level (core vs. context) based on the concept of **nuclearity in RST** [1].

0

hierarchically ordered sentences

2. Rhetorical Relation Identification:

CORE

CONTEX

- restores the **semantic relationship** between a pair of split sentences by identifying and classifying the rhetorical relation that holds between them.
- is based on syntactic and lexical features to extract cue phrases from the parse tree which are mapped to rhetorical relations using a predefined list of rhetorical cue words [2].

semantically interconnected sentences

"although" -> Contrast

CORE

CONTEXT

A fluoroscopic study ... is typically the next step in management.

If volvulus is suspected, caution with non water soluble contrast is mandatory as ... operative complications.

OUTPUT

[1] Mann and Thompson, 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. [2] Taboada and Das, 2013. Annotation upon Annotation: Adding Signalling Information to a Corpus of Discourse Relations. References [3] Sulem et al., 2018. Semantic Structural Evaluation for Text Simplification. [4] Niklaus et al., 2019. Transforming Complex Sentences into a Semantic Hierarchy. [5] Carlson et al., 2002. RST Discourse Treebank.